Report on

A Model of Saliency-Based Visual Attention for Rapid Scene Analysis

Jyothsna Shashikumar Sastry – 2571729
Department of Computer Science
Saarland University
Germany, Saarbrücken 66123
Email: s8jysast@stud.uni-saarland.de

***Abstract*— The term "scene analysis" refers to the process of interpreting a scene by identifying and inspecting the salient elements that constitute it. Detection of pedestrians and traffic sign elements in a scene are some important applications where rapid scene analysis is essential. Here, we discuss a model inspired by the architecture inherent in visual systems of early primates for rapid scene analysis. The proposed technique extracts feature maps at multiple scales and then combines them to produce a saliency map. This is input to a neural network which identifies salient locations in an efficient manner. The experimental results show that the proposed model, despite its simple architecture, performs better in comparison with other models for scene analysis.**

***Index Terms*— Feature Map, Neural Network, Rapid Scene Analysis, Saliency Map, Visual systems**

I. Introduction

HUMANS can analyze complex scenes despite the limited abilities of the neuronal hardware [1]. Studies conducted to understand visual attention suggest that humans reduce the complexity of scene analysis by selecting a subset of the information presented for further processing [2]. This selection is guided by the "focus of attention", a region of the visual field which scans the scene. A task-independent scan is rapid and controlled by salient locations in the scene. A task-dependent scan is rather slower and controlled by will [3].

The saliency based model of attention used in [1] is based on Human visual strategy explained by the Feature Integration Theory. This hypothesis suggests that attention must be directed serially to each stimulus in a display whenever conjunctions of more than one separable feature are needed to characterize or distinguish the possible objects presented [4]. Let us understand the visual process in humans for a better understanding of the aforementioned hypothesis. When presented with an image, the feature extraction happens at an early stage and the semantic information is gathered at a later stage. Hence, the visual system focuses attention to those locations constituted by separable features. The model in [1] is developed to mimic this behavior. The image is decomposed into a set of feature maps. The locations that stand out from their surroundings in each feature map are fed to a "saliency map" which guides the shift in attention. A neural network outputs the locations in the decreasing order of saliency. This highly parallel method is best suited for rapid selection of salient features in object recognition tasks. This can be demonstrated with the help of an activity as simple as taking a quick glance at a scene and identifying the salient locations.

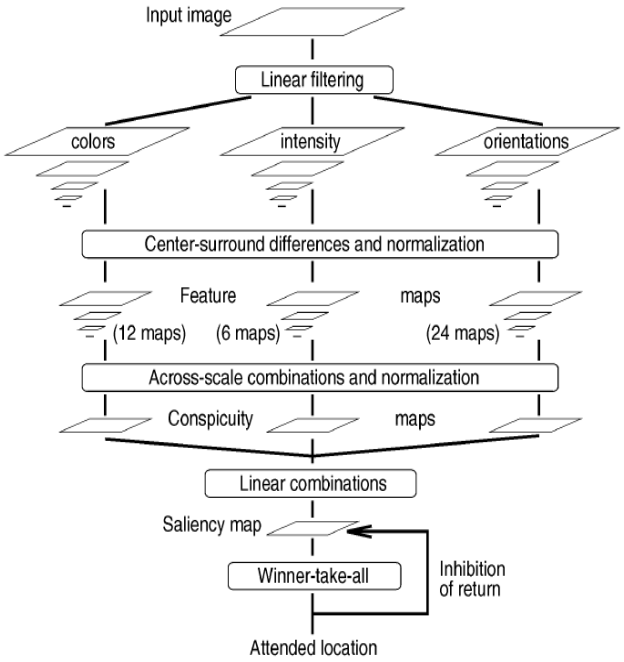

Figure 1: General architecture of the saliency based model in [1]

II. Architecture

This section explains the architecture of the Saliency based model presented in Figure 1. Static color images with a resolution of 640 × 480 are used as the input to the model. The model performs multi-scale feature extraction at nine spatial scales. Gaussian pyramids are used as they provide image representation at multiple scales by reducing the size of the image in every dimension by 2 as we move from a fine scale to a coarser scale. This is equivalent to successively performing low pass filtering on the subsampled input image at each level.

Figure 2 illustrates the working of Gaussian pyramids on an input image. If x×y is the resolution of the input image, the Gaussian pyramid at the coarse scale has resolution $\frac{x}{2} \times \frac{y}{2}$

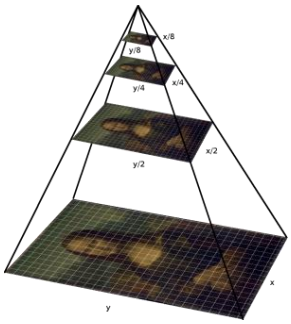

Figure 2: Illustration of Gaussian pyramids. Source:Google

Our eyes are sensitive to local spatial discontinuities and this helps to detect locations that stand out from their surroundings [5]. Feature extraction is performed using Center Surround Difference that mimics our visual system. The difference between two images at different scales is computed by interpolating the coarse image to the finer scale and performing pixel wise subtractions as shown in Figure 3. This across-scale difference between two maps is denoted by "Θ". The center is a pixel at scale c ∈ {2, 3, 4}. The corresponding pixel at scale s = c + δ, δ ∈ {3, 4} forms the surround. This yields a total of 6 pairs of center and surround combinations for multi-scale feature extraction.

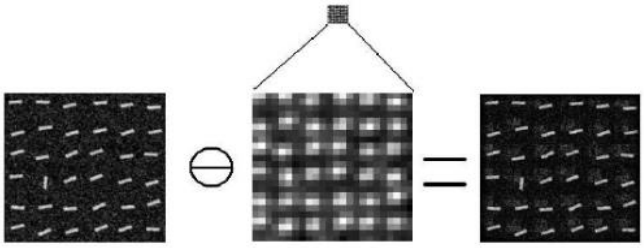

Figure 3: Illustration of across-scale center surround differences. Source:Google

A. Feature Extraction

The efficiency of a model for Rapid Scene analysis is dependent on the feature types the model is capable of extracting. Reference [1] implements the following feature maps using linear filtering, center surround differences and normalization:

1. Intensity contrast maps

An object in a scene is salient and easily detected when its intensity differs from that of its surroundings. Reference [5] attributes this to the neurons in the human visual system which are sensitive to bright centers with dark surrounds or dark centers with bright surrounds. The three channels in the input color image are Red(r), Green(g) and Blue(b). The measure of intensity at each pixel can be obtained by averaging over r, g and b values at that pixel.

$$I = (r + g + b) / 3$$

To aid multiscale feature extraction, Gaussian pyramids $I(\sigma)$ where $\sigma \in [0..8]$ are computed for the intensity map I at 9 scales. Regions of the image having low intensity can be ignored as the hue variations in these regions are not perceivable and hence, not salient. Therefore, [1] fixes a threshold for intensity normalization. The saliency based model chooses only those pixels with Intensity I > (1/10)th of I_{max} (maximum intensity over entire image) for normalization. The two kinds of sensitivities are captured in 6 maps obtained by computing center surround differences.

$$I(c, s) = |I(c) \Theta I(s)|$$

where $c \in \{2, 3, 4\}$ and $s = c + \delta$, with $\delta \in \{3, 4\}$.

2. Color contrast maps

Figure 4: Demonstration of chromatic opponency. Source:Google

The red apple and the yellow block stand out from their surroundings in Figure 4. Reference [6] based on Color tuning in Human visual cortex proposes that this is due to the chromatic opponency that exists between the pair of colors red-green and blue-yellow. The neurons at the center are excited by one color and inhibited by the other. The surrounding neurons, on the other hand, are excited by the opponent color and this makes it easy for the human eyes to distinguish the object from its surroundings [6]. On the basis of this proposal, the saliency based model in [1] creates four broadly tuned color channels R, G, B and Y

$$R = r - (g + b)/2$$
$$G = g - (r + b)/2$$
$$B = b - (r + g)/2$$
$$Y = (r + g)/2 - |r - g|/2 - b$$

and their Gaussian pyramids at 9 scales R(σ), G(σ), B(σ) and Y(σ). The center surround differences are computed according to the color opponency as follows

$$RG(c, s) = \left|\left(R(c) - G(c)\right) \Theta \left(G(s) - R(s)\right)\right|$$
$$BY(c, s) = \left|\left(B(c) - Y(c)\right) \Theta \left(Y(s) - B(s)\right)\right|$$

3. Orientation maps

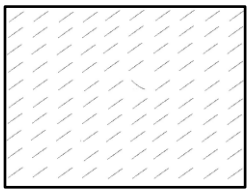

Figure 5: Illustration of breaks in pattern. Source:Google

When presented with an image as shown, our attention is drawn to that one location in the image where the line has an orientation different from its surroundings. Human visual attention system can identify breaks in patterns like in the Figure 5 with the aid of orientation sensitive neurons of the cortex [5]. Orientation information can be obtained by convolving the intensity image I with orientational filters. The intensity Image I is convolved with Gabor filters at nine scales [0...8] with Θ ∈ {0^0, 45^0, 90^0, 135^0} as the directions for filtering in [1]. Gabor filters are used as they show a strong response for image locations that have structures aligned with Θ

$$O(c, s, \Theta) = |O(c, \Theta) \Theta O(s, \Theta)|$$

B. *Saliency Map*

The above feature extraction procedure yields a total of 42 feature maps. 6 of these maps have intensity contrast information, 12 contain color contrast information and 24 of them contain orientation change information. These maps are combined to form a saliency map (SM) which assigns each location in the input image to a scalar that quantifies its saliency. The simplest way to obtain a single map with all the salient locations is to combine all 42 feature maps using across-scale additions. However, [1] advocates against this idea as the less salient objects present in large number of maps may overshadow the more-salient objects appearing only in a few maps. Also, the features are not always comparable to each other as they contain different scales and are extracted using different mechanisms. Reference [1] proposes a map normalization operator *N(.)*

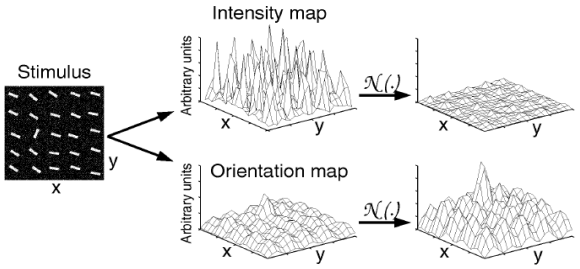

Figure 6: Map normalization operator *N(.)* in [1]

The following steps summarize the working of this operator.

1. Extract feature maps from the input image
2. Normalize each feature map to a fixed range [0…M]
3. Find location of global maximum M
4. Compute average $\overline{m}$ of the local maxima
5. Multiply map by (M-$\overline{m}$)^2

From Figure 6, if this difference is large, it indicates that the global maximum truly stands out from the local maxima. A small difference implies that the global maximum does not vary hugely from the local maxima and hence, the map contains nothing unique. This way, *N(.)* promotes maps with small number of conspicuous locations and suppresses maps containing homogenous locations.

The three conspicuity maps $\overline{I}$, $\overline{C}$ and $\overline{O}$ are formed using across scale additions at σ = 4 (each feature map is reduced to Gaussian scale 4 before performing point wise additions) [1]

$$\overline{I} = \oplus \oplus N\big(I(c,s)\big)$$

$$\overline{C} = \oplus \oplus \big[N\big(RG(c,s)\big) + N\big(BY(c,s)\big)\big]$$

$$\overline{O} = \sum_{\Theta \in \{0^0,45^0,90^0,135^0\}} N\Big(\oplus \oplus N\big(O(c,s,\Theta)\big)\Big)$$

where $c \in \{2, 3, 4\}$ and $s = c + \delta$, with $\delta \in \{3, 4\}$. The three conspicuity maps are normalized and added to obtain S as shown in the Figure 7.

$$S = \frac{1}{3}\Big(N\big(\overline{I}\big) + N\big(\overline{C}\big) + N\big(\overline{O}\big)\Big)$$

The Saliency Map (SM) is modeled as a 2D layer of leaky *integrate-and-fire neurons* at scale 4 [1]. They are called so because of their ability to integrate the input charges and fire when the accumulated charge exceeds a threshold. Each neuron in the network is activated by input from the *S* (map obtained on averaging the normalized conspicuity maps). The salient locations are identified using a "Winner-take-all" (WTA) neural network [1] whose neurons are excited by the corresponding SM neurons. The first WTA neuron whose accumulated charge exceeds the threshold is declared the "winner" and is said to trigger the following three responses

1. FOA shifts to the winner neuron
2. All the neurons in the WTA are reset
3. Only the SM neurons in FOA are reset

Steps 2 and 3 are referred to as global inhibition of the WTA neurons and local inhibition of the SM neurons respectively in [1]. Step 3 is performed to prevent the FOA from returning to the previously attended location and helps in dynamic shifting of the focus to the next most salient location.

In Figure 7, the neural network selected the telephone box which appeared strongly in $\overline{C}$ as the first salient location. The inhibition of return inhibits this location in SM and FOA shifts to unattended locations.

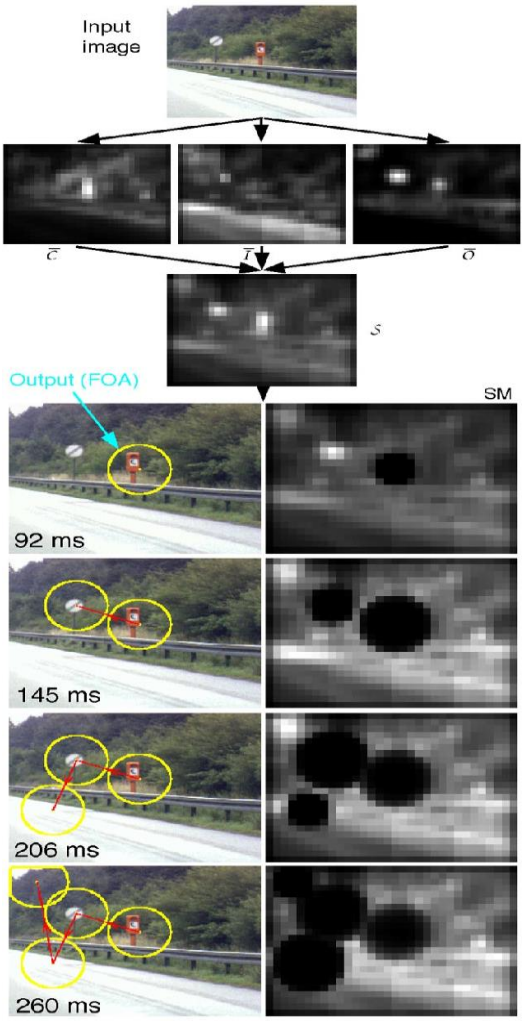

Figure 7: Operation of the model with natural image [1]

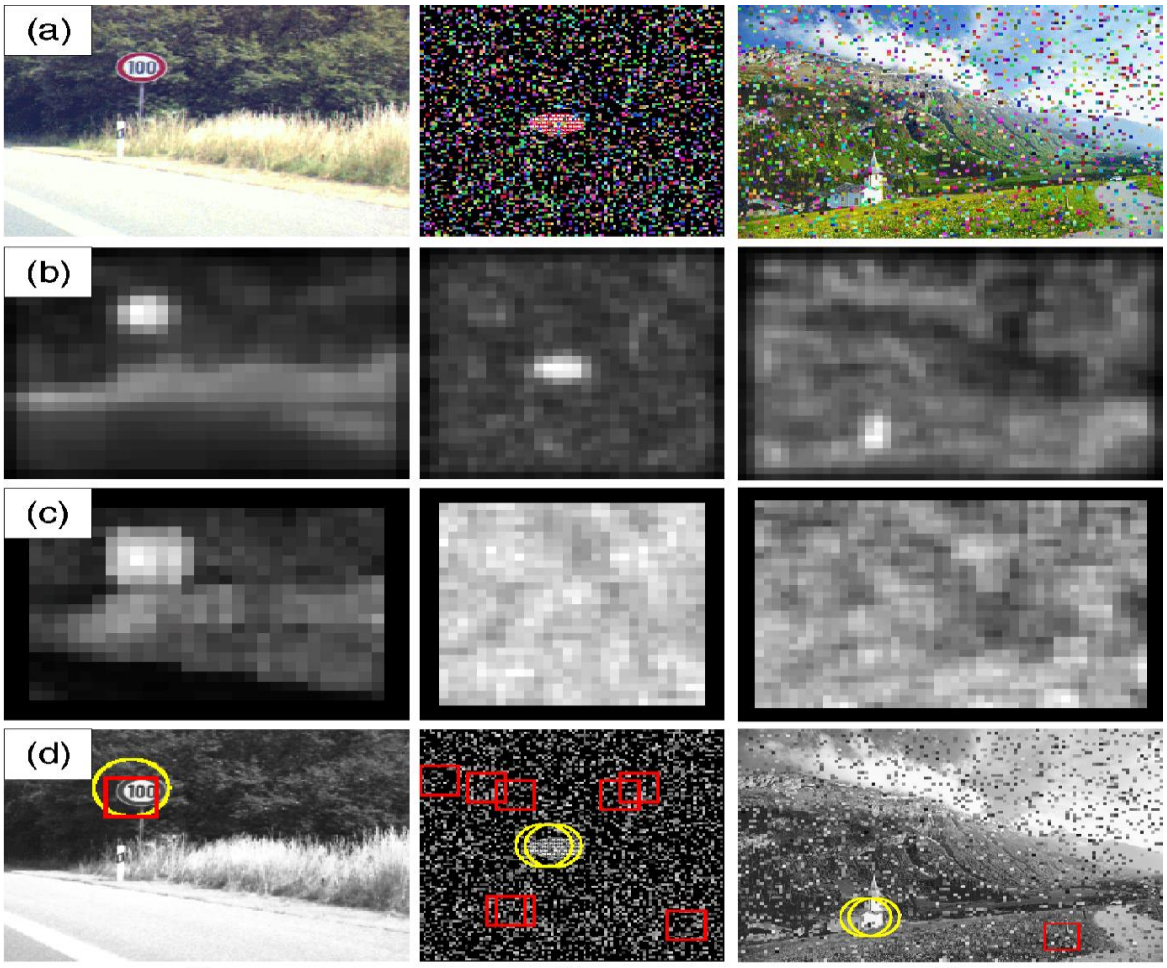

Figure 8: Comparison between the output of Saliency based model and SFC model in the presence of noise [1]

C. Comparison with Spatial Frequency Content Models

Reference [1] discusses the Spatial Frequency Content (SFC) model to compare its results with that of the saliency based approach. An eye tracking device was used to record the eye trajectory to analyze local spatial frequency distributions while viewing gray-scale images. The SFC was found to be higher at salient locations as opposed to the non-salient locations [7]. Reference [1] measures SFC as the average number of non-negligible coefficients after applying 2D Fast Fourier Transforms on 16×16 patches of $I(2)$, $R(2)$,$G(2)$,$B(2)$ and $Y(2)$ map. The scale and size are chosen such that the SFC model and saliency based model are equally sensitive to the changes in frequency and resolution.

The SFC map thus created is compared to the saliency map in Figure 8 for three example images in (a). (b) shows the maps input to SM and (c) shows the SFC maps. (d) shows the salient locations identified by saliency based model (yellow circle) and SFC model (red squares). It is evident from (d) that both models perform well in case of noiseless input and saliency maps are robust to noise while SFCs perform poorly when the input image is noisy.

III. RESULTS

Several tests have been performed to ensure proper functioning and robustness of the model in [1]. The saliency based model was tested using several objects of the same shape but different contrast with respect to the background. It was found that the model attended to the objects in the decreasing order of contrast. The model was also tested with real images like paintings and landscapes. The model attended to all the objects of interest. The model was successful in reproducing human performance for pop-out tasks when the target differed from its surroundings by color, intensity or orientation as in Figure 9 [4]. Reference [1] concludes that the model is robust to noise when the added noise does not conflict with the main feature of the object. In Figure 10, the input is a color image of resolution 768×512. The two people that form the salient objects are unique because of the contrasting color with the surroundings. The performance was tested by introducing white noise with different densities. The model succeeded in identifying the two people as salient. The same test was repeated with the introduction of colored noise. The model

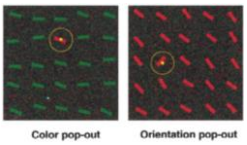

Figure 9: Illustration of pop out in [1]

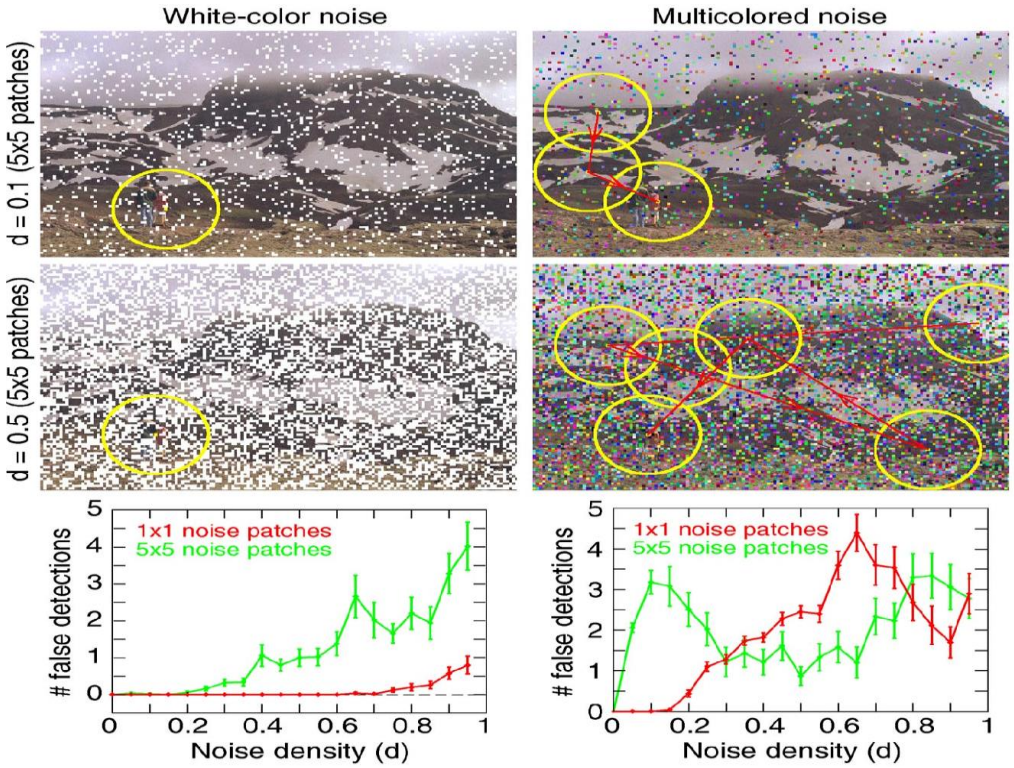

Figure 10: Illustration of the influence of noise on detection performance in [1]

identified various salient locations in addition to the uniquely colored target because the noise has similar properties as the target (color). Hence, the model attended to various locations salient for other features (intensity).

The graph depicts that, in the presence of noise, the model results in false positives with the increase in noise density. In case of white noise with colored target, a 1×1 noise patch is more stable and false detections increase only for large noise densities. 5×5 noise patches have false detections for noise density as low as 0.2. In case of colored noise with colored target, the number of false positives increase alarmingly at high densities for a 1×1 noise patch and 5×5 noise patches can result in false detections with the introduction of very little noise.

A. *Strengths and Limitations*

A saliency based model with adequate feature maps can be used to guide bottom up attention [8]. The model can be deployed to perform tasks in real time, owing to the highly parallel implementation of feature extraction. Despite the strong performance with natural scenes, the model does have some limitations. The result of pop out tasks depends on the feature types implemented. The model in [1] will fail to identify important targets like T-junctions, line endings and other unimplemented feature types. The model fails to perform tasks like contour completion and closure due to the absence of recurrence mechanisms in the feature maps. The absence of the magnocellular motion channel makes the model unfit to be used for applications like optical character recognition.

IV. Conclusion

The saliency based model is conceptually simple and can reproduce the performance of visual system in primates. Despite its simple architecture, it is highly relevant for modern day tasks that involve rapid scene analysis. It can be applied in modern day use cases like autonomous driving and robotics that rely on real time scene understanding for decision making. The model can also be extended to suit specific target detection tasks by implementing feature types that constitute the target of interest.

References.

[1] Laurent Itti, Christof Koch, and Ernst Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 11, NOVEMBER 1998

[2] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, and F. Nuflo, "Modelling Visual Attention via Selective Tuning,"Artificial Intelligence, vol. 78, no. 1-2, pp. 507–545, Oct. 1995.

[3] E. Niebur and C. Koch, "Computational Architectures for Attention," R. Parasuraman, ed., The Attentive Brain, pp. 163–186. Cambridge,Mass.: MIT Press, 1998

[4] A.M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," Cognitive Psychology, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[5] A.G. Leventhal, The Neural Basis of Visual Function: Vision and Visual Dysfunction, vol. 4. Boca Raton, Fla.: CRC Press, 1991.

[6] S. Engel, X. Zhang, and B. Wandell, "Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging," Nature, vol. 388, no. 6,637, pp. 68–71, July 1997.

[7] P. Reinagel and A.M. Zador, "The Effect of Gaze on Natural Scene Statistics," Neural Information and Coding Workshop, Snowbird, Utah, 16-20 Mar. 1997.

[8] J.P. Gottlieb, M. Kusunoki, and M.E. Goldberg, "The Representation of Visual Salience in Monkey Parietal Cortex," Nature, vol. 391, no. 6,666, pp. 481-484, Jan. 1998.